Introduction

Throughout this course, we shall be interested in the analog to digital conversion of signals $f(t),\ t \in \mathbb{R}$. We shall always assume $f \in L_2$ and usually assume additional properties of $f$ in order to get meaningful results. In particular, we want to study two mappings: the encoding of $f$ into bit streams and the decoding of the bit streams into approximations or estimates of $f$,

**Equation:**

$$\begin{aligned} E: &\quad f \to \text{bits streams} \quad (\text{Encoder}) \\ D: &\quad \text{bits streams} \to f \quad (\text{Decoder}) \end{aligned}$$

where $\hat{f}$ is the approximation of $f$ defined by $\hat{f} := D(E(f))$. In general, $\hat{f} \neq f$, so we shall need some way of quantifying how well $\hat{f}$ approximates $f$. Normally, the distortion between is measured by some norm $\| f - \hat{f} \|$. Typical choices include:

**Equation:**

$$\begin{aligned} \text{the } L_2 \text{ norm} \quad \| f \|_{L_2} &:= \left( \int |f(t)|^2 \, dt \right)^{1/2} \\ \text{the } L_\infty \text{ norm} \quad \| f \|_{L_\infty} &:= \sup_t |f(t)| \\ \text{the } L_p \text{ norm} \quad \| f \|_{L_p} &:= \left( \int |f(t)|^p \, dt \right)^{1/p} \end{aligned}$$

## The Shannon-Whitaker Sampling Theorem

The classical theory behind the encoding analog signals into bit streams and decoding bit streams back into signals, rests on a famous sampling theorem which is typically refereed to as the Shannon-Whitaker Sampling Theorem. In this course, this sampling theory will serve as a benchmark to which we shall compare the new theory of compressed sensing.

To introduce the Shannon-Whitaker theory, we first define the class of bandlimited signals. A bandlimited signal is a signal whose Fourier transform only has finite support. We shall denote this class as $B_A$ and define it in the following way:
**Equation:**

$$B_A := \{f \in L_2(\mathbb{R}) : \widehat{f}(\omega) = 0,\ |\omega| \geq A\pi\}.$$

Here, the Fourier transform of $f$ is defined by
**Equation:**

$$\widehat{f}(\omega) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t) e^{-i\omega t}\, dt.$$

This formula holds for any $f \in L_1$ and extends easily to $f \in L_2$ via limits. The inversion of the Fourier transform is given by
**Equation:**

$$f(t) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}(\omega) e^{i\omega t}\, d\omega.$$

## Shannon-Whitaker Sampling Theorem

If $f \in B_A$, then $f$ can be uniquely determined by the uniformly spaced samples $f\left(\frac{n}{A}\right)$ and in fact, is given by
**Equation:**

$$f\left(t\right)=\sum_{n\in\mathbb{Z}}f\Big(\frac{n}{A}\Big)\ \operatorname{sinc}\left(\pi(\mathrm{A}t-\mathrm{n})\right),$$

where $\operatorname{sinc}\left(\mathrm{t}\right)=\dfrac{\sin\,\mathrm{t}}{\mathrm{t}}$.

It is enough to consider $A=1$, since all other cases can be reduced to this through a simple change of variables. Because $f\in B_{A=1}$, the Fourier inversion formula takes the form
**Equation:**

$$f\left(t\right)=\frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi}\widehat{f}\left(\omega\right)e^{i\omega t}\ d\omega.$$

Define $F(\omega)$ as the $2\pi$ periodization of $\widehat{f}$,
**Equation:**

$$F\left(\omega\right):=\sum_{n\in\mathbb{Z}}\widehat{f}\left(\omega-2n\pi\right).$$

Because $F(\omega)$ is periodic, it admits a Fourier series representation
**Equation:**

$$F\left(\omega\right)=\sum_{n\in\mathbb{Z}}c_{n}e^{-in\omega},$$

where the Fourier coefficients $c_n$ given by
**Equation:**

$$\begin{aligned}c_{n}\ &=\frac{1}{2\pi}\int_{-\pi}^{\pi}F\left(\omega\right)e^{in\omega}\ d\omega\\&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\widehat{f}\left(\omega\right)e^{in\omega}\ d\omega.\end{aligned}$$

By comparing ([link]) with ([link]), we conclude that
**Equation:**

$$c_n = \frac{1}{\sqrt{2\pi}} f\left(n\right).$$

Therefore by plugging ([link]) back into ([link]), we have that
**Equation:**

$$F\left(\omega\right) = \frac{1}{\sqrt{2\pi}} \sum_{n \in \mathbb{Z}} f\left(n\right) e^{-in\omega}.$$

Now, because
**Equation:**

$$\widehat{f}\left(\omega\right) = F\left(\omega\right) \chi_{[-\pi,\pi]} = \frac{1}{\sqrt{2\pi}} \sum_{n \in \mathbb{Z}} f\left(n\right) e^{-in\omega} \chi_{[-\pi,\pi]},$$

and because of the facts that
**Equation:**

$$\begin{aligned} \mathscr{F}\left(\chi_{[-\pi,\pi]}\right) &= \frac{1}{\sqrt{2\pi}} \operatorname{sinc}\left(\pi\omega\right) \quad \text{and} \\ \mathscr{F}(g(t-n)) &= e^{-in\omega} \mathscr{F}(g(t)), \end{aligned}$$

we conclude
**Equation:**

$$f\left(t\right) = \sum_{n \in \mathbb{Z}} f(n) \ \operatorname{sinc}\left(\pi(\mathrm{t}-\mathrm{n})\right).$$

Comments:

1. (Good news) The set $\{\text{sinc}\left(\pi(\text{t}-\text{n})\right)\}_{\text{n}\in\mathbb{Z}}$ is an orthogonal system and therefore, has the property that the $L_2$ norm of the function and its Fourier coefficients are related by,
   **Equation:**

$$\parallel f \parallel_{L_2}^2 = 2\pi \sum_{n\in\mathbb{Z}} \left| f\left(n\right) \right|^2$$

2. (Bad news) The representation of $f$ in terms of sinc functions is not a stable representation, i.e.
   **Equation:**

$$\sum_{n\in\mathbb{Z}} |\text{sinc}\left(\pi(\text{t}-\text{n})\right)| \approx \sum_{\text{n}\in\mathbb{Z}} \frac{1}{|\text{t}-\text{n}|+1} \rightarrow \text{divergences}$$

Optimal Encoding

We shall consider now the encoding of signals on $[-T, T]$ where $T > 0$ is fixed. Ultimately we shall be interested in encoding classes of bandlimited signals like the class $B_A$ However, we begin the story by considering the more general setting of encoding the elements of any given compact subset $K$ of a normed linear space $X$. One can determine the best encoding of $K$ by what is known as the Kolmogorov entropy of $K$ in $X$.

To begin, let us consider an encoder-decoder pair $(E, D)$ $E$ maps $K$ to a finite stream of bits. $D$ maps a stream of bits to a signal in $X$. This is illustrated in [link]. Note that many functions can be mapped onto the same bitstream.

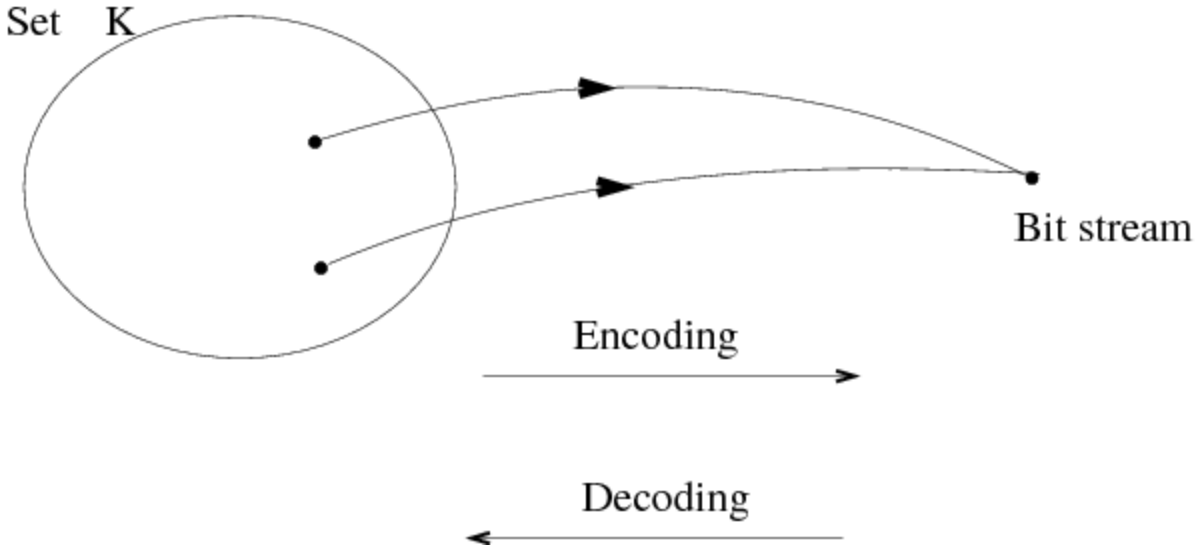



Illustration of encoding and decoding.

Define the distortion $d$ for this encoder-decoder by
**Equation:**

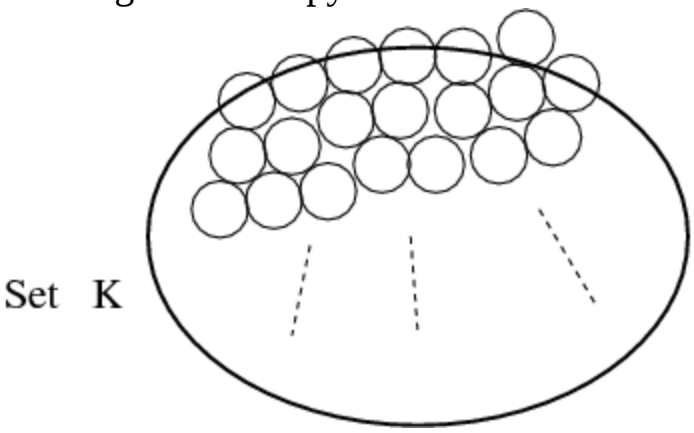
$$d(K, E, D, X) := \sup_{f \in K} \| f - D(Ef) \|_X.$$

Let $n(K, E) = \sup_{f \in K} \#Ef$ where $\#Ef$ is the number of bits in the bitstream $Ef$. Thus $n$ is the maximum length of the bitstreams for the various $f \in K$. There are two ways we can define optimal encoding:

1. Prescribe $\epsilon$, the maximum distortion that we are willing to tolerate. For this $\epsilon$, find the smallest $n_\epsilon(K, X) := \inf_{(E,D)} \{n(K, E) : d(K, E, D, X) \leq \epsilon\}$. This is the smallest bit budget under which we could encode all elements of $K$ to distortion $\epsilon$.
2. Prescribe $N$ : find the smallest distortion $d(K, E, D, X)$ over all $E, D$ with $n(K, E) \leq N$. This is the best encoding performance possible with a prescribed bit budget.

There is a simple mathematical solution to these two encoding problems based on the notion of Kolmogorov Entropy.

Kolmogorov Entropy




Coverings of $K$ by balls of radius $\epsilon$.

Given $\epsilon > 0$, and the compact set $K$, consider all coverings of $K$ by balls of radius $\epsilon$, as shown in [link]. In other words,

**Equation:**

$$K \subseteq U_{i=1}^{N} b(f_i, \epsilon).$$

Let $N_\epsilon := \inf \ \{N : \text{over all such covers}\}$. $N_\epsilon(K)$ is called the covering number of $K$. Since it depends on $X$ and $K$, we write it as $N_\epsilon = N_\epsilon(K, X)$.

**Definition**

Kolmogorov entropy

The Kolmogorov entropy, denoted by $H_\epsilon(K, X)$, of the compact set $K$ in $X$ is defined as the logarithm of the covering number:

**Equation:**

$$H_\epsilon(K, X) = \log N_\epsilon(K, X).$$

The Kolmogorov entropy solves our problem of optimal encoding in the sense of the following theorem.

For any compact set $K \subset X$, we have $n_\epsilon(K, X) = \lceil H_\epsilon(K, X) \rceil$, where $\lceil \cdot \rceil$ is the ceiling function.

Sketch: We can define an encoder-decoder as follows To encode: Say $f \in K$. Just specify which ball it is covered by. Because the number of balls is $N_\epsilon(K, X$ , we need at most $\log N_\epsilon(K, X)$ bits to specify any such ball ball.

To decode: Just take the center of the ball specified by the bitstream.

It is now easy to see that this encoder-decoder pair is optimal in either of the senses given above.

The above encoder is not practical. However, the Kolmogorov entropy tells us the best performance we can expect from any encoder-decoder pair. Kolmogorov entropy is defined in the deterministic setting. It is the analogue of the Shannon entropy which is defined in a stochastic setting.

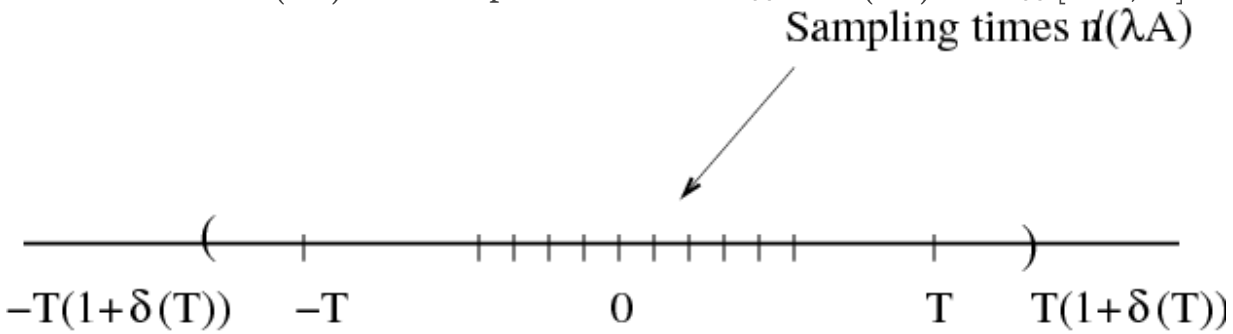
Optimal Encoding of Bandlimited Signals

We now turn back to the encoding of signals. We are interested in encoding the set

**Equation:**

$$B_A(M) = \{f \in B_A : |f(t)| \leq M, t \in \mathbb{R}\}$$

where $M$ is arbitrary but fixed. We shall restrict our discussion to the case where distortion is measured in $L_\infty[-T, T]$ where $T > 0$ is arbitrary but fixed. Then, $B_A(M)$ is a compact subset of $L_\infty$: $B_A(M) \subseteq L_\infty[-T, T]$.




Sample points $\frac{n}{\lambda A}$ are chosen in the interval $[-T(1+\delta), T(1+\delta)]$.

We shall sketch how one can construct an asymptotically optimal encoder/decoder for $B_A$. The details for this construction can be found in [link].

We know $\widehat{f}(\omega) = 0$ for $|\omega| \geq A\pi$, and $|f| \leq M$. How can we encode $f$ in practice? We begin by chosing $\lambda = \lambda(T) > 1$ (see [link]) which will represent a slight oversampling factor we shall utilize. Given a target distortion $\epsilon > 0$, we choose $k$ so that $2^{-k-1} < \epsilon \leq 2^{-k}$. Given $f$, we shall encode $f$ by first taking samples $f\left(\frac{n}{\lambda A}\right)$ for $\frac{n}{\lambda A} \in [-T(1+\delta), T(1+\delta)]$ where $\delta(T) > 0$. In other words, we sample $f$ on a slightly larger interval than $[-T, T]$. For each sample $f\left(\frac{n}{\lambda A}\right)$, we shall use the first $k + k_0(T)$ bits of its binary expansion. In other words, our encoder takes $f$ and the

samples $f\left(\frac{n}{\lambda A}\right)$ and then assigns to $f\left(\frac{n}{\lambda A}\right)$ the first $k+k_0\left(T\right)$ bits of this number.

To decode, the receiver would take the bits and construct the approximation $f\left(\frac{n}{\lambda A}\right)$ to $f\left(\frac{n}{A\lambda}\right)$ from the bits provided. Notice that we have the accuracy
**Equation:**

$$f(\frac{n}{\lambda A})-f\left(\frac{n}{\lambda A}\right) \leq 2^{-k-k_0}\cdot M.$$

We utilize the function $g_\lambda$ satisfying ([link]) to define
**Equation:**

$$f\ (t)=\sum_{n\in N_T} f\left(\frac{n}{\lambda A}\right)g_\lambda\left(\lambda A t-n\right),$$

where
**Equation:**

$$N_T:=\{n:\ -T(1+\delta)\leq\frac{n}{\lambda A}\leq T(1+\delta)\}.$$

We then have
**Equation:**

$$\begin{aligned}|f(t)-f\ (t)| \ \leq\ & \sum_{n\in N_T} f\left(\frac{n}{\lambda A}\right)-f\left(\frac{n}{\lambda A}\right)\cdot|g_\lambda\left(\lambda A t-n\right)| \\ & +\sum_{|\frac{n}{\lambda A}|>T(1+\delta)} f\left(\frac{n}{\lambda A}\right)\cdot|g_\lambda\left(\lambda A t-n\right)|\end{aligned}$$

The term $f\left(\frac{n}{\lambda A}\right) - f\left(\frac{n}{\lambda A}\right)$ that appears in the first summation in ([link]) is bounded by $M \cdot 2^{-k-k_0}$. The term $f\left(\frac{n}{\lambda A}\right)$ that appears in the second summation in the same equation is bounded by $M$. Therefore,
**Equation:**

$$\begin{aligned} |f(t) - f\ (t)| \ &\leq \sum_{n \in N_T} M \cdot 2^{-k-k_0} \cdot |g_\lambda\left(\lambda A t - n\right)| \\ &\quad + \sum_{|\frac{n}{\lambda A}| > T(1+\delta)} M \cdot |g_\lambda\left(\lambda A t - n\right)| =: S_1 + S_2 \end{aligned}$$

We can estimate $S_1$ by
**Equation:**

$$\begin{aligned} S_1 \ &= \sum_{n \in N_T} M \cdot 2^{-k-k_0} \cdot |g_\lambda\left(\lambda A t - n\right)| \\ &\leq M \cdot 2^{-k-k_0} \cdot \sum_n \ g_\lambda\left(\lambda A t - n\right)| \\ &\leq M \cdot C_0\left(\lambda\right) \cdot 2^{-k-k_0} \qquad \left(\text{because } g(\cdot) \text{ decays fast}\right) \end{aligned}$$

Therefore, if we choose $k_0$ sufficiently large, then
$S_1 \leq M \cdot C_0\left(\lambda\right) \cdot 2^{-k-k_0} \leq \frac{\epsilon}{2}$. The second summation $S_2$ can also be bounded by $\epsilon/2$ by using the fast decay of the function $g_\lambda$ (see ([link])).

To make the encoder/decoder specific we need to precisely define $\delta$ and $\lambda$. It turns out that the best choices (in terms of bit rate performance on the class $B_A$) depend on $T$. But $\delta_T \to 0$ and $\lambda_T \to 1$ as $T \to \infty$. Recall that Shannon sampling requires $2T\lambda A$ samples. Since our encoder/decoder uses $k + k_0$ bits per sample, the total number of bits is $(k + k_0) \cdot 2\lambda A T(1 + \delta)$, and so coding will require roughly $k$ bits per Shannon sample.

This encoder/decoder can be proven to be optimal in the sense of averaged performance as we shall now describe. The average of performance of optimal encoding is defined by

**Equation:**

$$\lim_{T \to \infty} \frac{n_\epsilon(B_A(M), L_\infty \lfloor -T, T \rfloor)}{2T}$$

If we replace the optimal bit rate $n_\epsilon$ in ([link]) by the number of bits required by our encoder/decoder then the resulting limit will be the same as that in ([link]).

In summary, to encode band limited signals on an interval $[-T, T]$, an optimal strategy is to sample at a slightly higher rate than Nyquist and on a slightly large interval than $[-T, T]$. Each sample should then be quantized by using the binary expansion of the sample. In this way, for an investment of $k$ bits per Nyquist rate sample, we get a distortion of $2^{-k}$.

To get a feel for the number of bits required by such an encoder, let us say $A = 10^6$ (signals band limited to 1Mhz). Say $T = 24$ hours $\approx 10^5$ seconds, and $k = 10$ bits. Then, $A \cdot k \cdot 2T = 10^6 \cdot 10 \cdot 10^5 = 10^{12}$ bits. This is too BIG!

The above encoding is is known as Pulse Coded Modulation (PCM). In practice, people frequently use another encoder called Sigma-Delta Modulation. Instead of oversampling just slightly, Sigma Delta over samples a lot and then assign only one (or a few) bits per sample.

Why is Sigma-Delta preferred to PCM in practice? There are two reasons commonly given:

1. Getting accurate samples, quantization, etc. is not practical because of noise. For better accuracy, we need more expensive hardware.
2. Noise shaping. In Sigma-Delta, the distortion is higher but the distortion is spread over frequencies outside of the desired range.

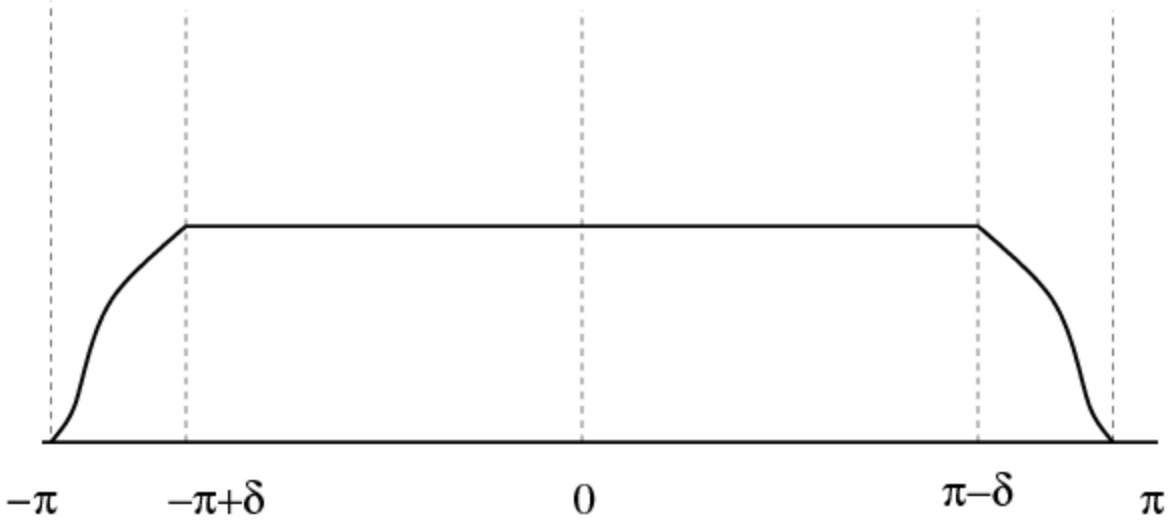
In PCM, the distortion decays exponentially (like $2^{-k}$), whereas for Sigma-Delta, the distortion decays like a polynomial (like $\frac{1}{k^m}$). Although the distortion decays faster in PCM, the distortion in Sigma-Delta is spread outside the desired frequency range.

Stable Signal Representations

To fix the instability of the Shannon representation, we assume that the signal is slightly more bandlimited than before
**Equation:**

$$\widehat{f}(\omega) = 0 \quad \text{for} \quad \left|\omega\right| \geq \pi - \delta,\ \delta > 0,$$

and instead of using $\chi_{[-\pi,\pi]}$, we multiply by another function $\hat{g}(\omega)$ which is very similar in form to the characteristic function, but decays at its boundaries in a smoother fashion (i.e. it has more derivatives). A candidate function $\hat{g}$ is sketched in [link].




Sketch of $\hat{g}$.

Now, it is a property of the Fourier transform that an increased smoothness in one domain translates into a faster decay in the other. Thus, we can fix our instability problem, by choosing $\hat{g}$ so that $\hat{g}$ is smooth and $\hat{g}(\omega) = 1$, $|\omega| \leq \pi - \delta$ and $\hat{g} = 0$, $|\omega| > \pi$. By choosing the smoothness of $g$ suitably large, we can, for any given $m \geq 1$, choose $g$ to satisfy
**Equation:**

$$\left|g\left(t\right)\right| \leq \frac{C}{\left(|t|+1\right)^{m}}$$

for some constant $C > 0$.

Using such a $\hat{g}$, we can rewrite ([link]) as
**Equation:**

$$\widehat{f}\left(\omega\right) = F\left(\omega\right)\hat{g}\left(\omega\right) = \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}} f\left(n\right)e^{-in\omega}\hat{g}\left(\omega\right).$$

Thus, we have the new representation
**Equation:**

$$f\left(t\right) = \sum_{n\in\mathbb{Z}} f\left(n\right)\ g\left(t-n\right),$$

where we gain stability from our additional assumption that the signal is bandlimited on $[-\pi-\delta, \pi-\delta]$.

Does this assumption really hurt? No, not really because if our signal is really bandlimited to $[-\pi, \pi]$ and not $[-\pi-\delta, \pi-\delta]$, we can always take a slightly larger bandwidth, say $[-\lambda\pi, \lambda\pi]$ where $\lambda$ is a little larger than one, and carry out the same analysis as above. Doing so, would only mean slightly oversampling the signal (small cost).

Recall that in the end we want to convert analog signals into bit streams. Thus far, we have the two representations
**Equation:**

$$\begin{aligned} f(t) &= \sum_{n\in\mathbb{Z}} f(n)\ \ \mathrm{sinc}\left(\pi(\mathrm{t}-\mathrm{n})\right), \\ f(t) &= \sum_{n\in\mathbb{Z}} f\Big(\frac{n}{\lambda}\Big)\ g\left(\lambda t-n\right). \end{aligned}$$

Shannon's Theorem tells us that if $f \in B_A$, we should sample $f$ at the Nyquist rate $A$ (which is twice the support of $\widehat{f}$) and then take the binary representation of the samples. Our more stable representation says to slightly oversample $f$ and then convert to a binary representation. Both representations offer perfect reconstruction, although in the more stable representation, one is straddled with the additional task of choosing an appropriate $\lambda$.

In practical situations, we shall be interested in approximating $f$ on an interval $[-T, T]$ for some $T > 0$ and not for all time. Questions we still want to answer include

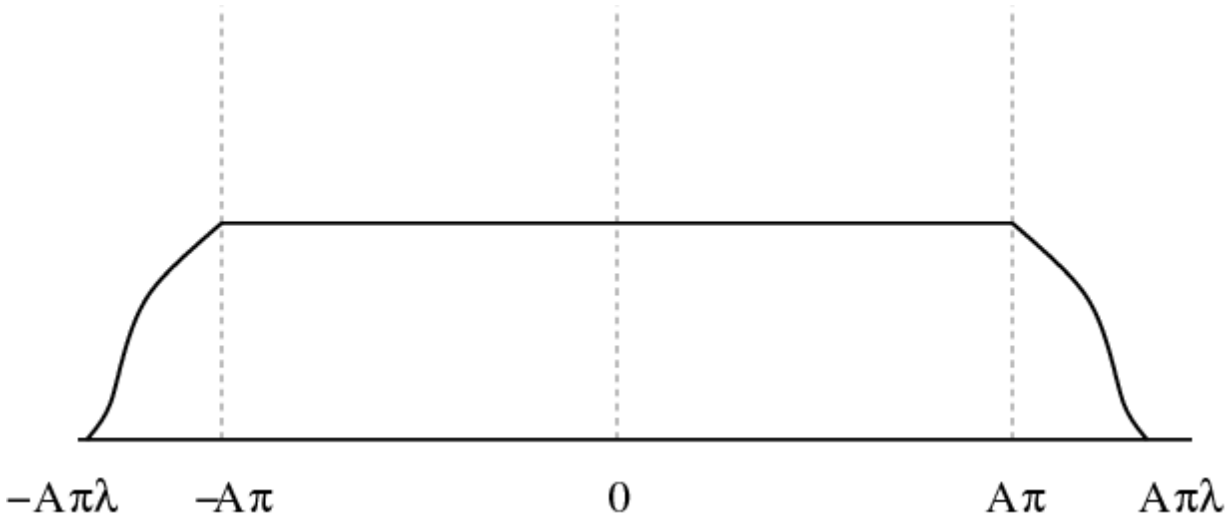
1. How many bits do we need to represent $f$ in $B_{A=1}$ on some interval $[-T, T]$ in the norm $L_\infty [-T, T]$?
2. Using this methodology, what is the optimal way of encoding?
3. How is the optimal encoding implemented?

Towards this end, we define
**Equation:**

$$B_A := \{f \in L_2 (\mathbb{R}) : |\widehat{f}(\omega)| = 0, |\omega| \geq A\pi\}.$$

Then for any $f \in B_A$, we can write
**Equation:**

$$f = \sum_n f\Big(\frac{n}{A}\Big) \cdot \text{sinc } \pi (At - n).$$




Fourier transform of $g_\lambda(\cdot)$.

In other words, samples at 0, $\pm\frac{1}{A}$, $\pm\frac{2}{A}$, $\cdots$ are sufficient to reconstruct $f$. Recall also that sinc $(x) = \frac{\sin(x)}{x}$ decays poorly (leading to numerical instability). We can overcome this problem by slight over-sampling. Say we over-sample by a factor $\lambda > 1$. Then, we can write

**Equation:**

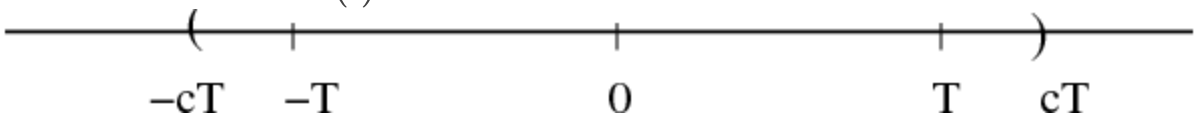
$$f = \sum f\Big(\frac{n}{\lambda A}\Big) g_\lambda(\lambda A t - n).$$

Hence we need samples at 0, $\pm\frac{1}{\lambda A}$, $\pm\frac{2}{\lambda A}$, etc. What is the advantage? Sampling more often than necessary buys us stability because we now have a choice for $g_\lambda(\cdot)$. If we choose $g_\lambda(\cdot)$ infinitely differentiable whose Fourier transform looks as shown in [link] we can obtain

**Equation:**

$$\Big| g_\lambda(t)| \leq \frac{c_{\lambda,k}}{(1+|t|)^k}, \quad k = 1, 2, \ldots$$

and therefore $g_\lambda(\cdot)$ decays very fast. In other words, a sample's influence is felt only locally. Note however, that over-sampling generates basis functions that are redundant (linearly dependent), unlike the integer translates of the $\mathrm{sinc}(\cdot)$ function.




To reconstruct signals in $[-T, T]$, the sampling interval is $[-cT, cT]$
.

If we restrict our reconstruction to $t$ in the interval $[-T, T]$, we will only need samples only from $[-cT, cT]$, for $c > 1$ (see [link]), because the distant samples will have little effect on the reconstruction in $[-T, T]$.

Preliminaries

We previously described Shannon's Theorem plus encoding: the Nyquist sampling rate is the minimal required sampling rate to recover the entire class of bandlimited signals. We have seen that this sampling rate may be prohibitively large for broadband signals. We see a way to improve upon this situation: we will pose a different model for the signals which is more restrictive than the assumption that the signals are bandlimited. Fortunately, there are several real world scenarios in which one knows much more information about the signals of interest. For example, they may be written in terms of very few fundamental building blocks (such as sine waves or chirps). This leads us to define new signal classes based on notions of sparsity and seek to determine if we can improve on sampling and encoding in this new setting.

Let us define the general setting for this section. Let $\mathbf{X}$ be a Banach space of functions. The typical examples are $\mathbf{X} = L_p(\mathbb{R}), L_p\left(\mathbb{R}^d\right), L_p(-T, T)$, $1 \leq p \leq \infty$. We denote the norm on $X$ by $\| \cdot \|_{\mathbf{X}}$. We define a dictionary $\mathscr{D}$ as any collection of functions $\mathscr{D} \subseteq \mathbf{X}$ such that $\| g \|_{\mathbf{X}} = 1$ for all $g \in \mathscr{D}$, i.e. all the elements of the dictionary are normalized. While the definition is very broad, in practice dictionaries usually have more structure. Some examples include $\mathscr{D} = B$, a basis for $\mathbf{X}$, such as (i) the Fourier basis on $[-\pi, \pi]$, (ii) a wavelet basis,[footnote] (iii) redundant families of waveforms of the form $\psi_{a,b,\sigma} = e^{-a(t-b)^2} e^{i\sigma x}$, i.e. $\mathscr{D} = \{\psi_{a,b,\sigma}\}_{a,b,\sigma}$, and (iv) wavelet packets.
Wavelet basis form orthonormal systems for $L_2(I)$.

**Definition**

We define the class of $n$-sparse signals as $\Sigma_n := \Sigma_n(\mathscr{D}) = \{s = \sum_{g \in \Lambda} c_g g, \Lambda \subseteq \mathscr{D}, \sharp\Lambda \leq n\}$. We also say that $s$ has sparsity $n$ in $\mathscr{D}$ if $s \in \Sigma_n(\mathscr{D})$, i.e. if it can be written as the linear combination of $n$ functions from $\mathscr{D}$. We note that $\Sigma_n$ is not a linear space; we instead have $\Sigma_n + \Sigma_n \subseteq \Sigma_{2n}$.

## New Signal Models

We now wish to consider new model classes for signals. Towards this end, let $\{\psi_j\}_{j=1}^{\infty}$ be an orthonormal basis for $L_2(-T,T)$. Thus for $f \in L_2$ we can write $f = \sum_{j=1}^{\infty} c_j(f)\psi_j$ where $(c_j(f)) \in \ell_2$. We will now build an encoder and decoder and analyze its performance on compact sets $K$. For example, we might want to encode signals in the space

$$X_p = \{f: (c_j(f)) \in \ell_p \}, \, 0 \leq p \leq 2$$

with norm

$$\| f \|_{X_p} := \| \left(c_j(f)\right) \|_{\ell_p}.$$

However, in this space the unit ball, $U(X_p)$ is not compact. To get a compact set we need more structure on the sequence $(c_j)$. Hence we define

$$Y^{\alpha} := \{f: \mid c_n(f) \mid \leq n^{-\alpha}, n = 1,2,\ldots\}$$

and we define the norm in this space as $\| f \|_{Y^{\alpha}} :=$ the smallest $c$ such that this holds. We now take

$$K = U(X_p) \cap U(Y^{\alpha})$$

to get a compact set. Notice that when $\alpha > 0$ is small the requirement for membership in $Y^{\alpha}$ is very mild.

Next, suppose that we choose a target distortion level $\varepsilon = 2^{-m}$. Given $f$, let

$$\Lambda_k := \Lambda_k(f) = \{j \in \{0,\ldots,N\}: 2^{-k-1} \leq \mid c_j(f) \mid < 2^{-k}\}$$

for $0 \leq k \leq M$, where $M := \lceil \frac{2m}{2-p} \rceil$. We then choose $N$ as the smallest integer so that

$$N^{-\alpha} \leq 2^{-M}$$

and thus

$$\log\ N \leq Cm.$$

It follows from the requirement that $f \in Y^{\alpha}$ that $\Lambda_k \subset \{1, \ldots, N\}$ for each $0 \leq k \leq M$.

Recall that

$$\#\Lambda_k\ 2^{(-k-1)p} \leq \sum_{c_j \in \Lambda_k} |c_j|^p \leq \parallel f \parallel_{X_p}^p.$$

Since $f \in U(X_p) \cap U(Y^{\alpha})$,

$$\#\Lambda_k \leq \parallel f \parallel_{\ell_p} 2^{(k+1)p} \leq 2^{(k+1)p}.$$

Hence, the total number of indices in all of the $\Lambda_k$, $0 \leq k \leq M$, is $O(2^{Mp})$.

To encode, for each $f$, we can send the following bits:

- Send $\log\ n$ bits to identify each index in $\Lambda_k$, for $0 \leq k \leq M$. This will require a total of $O(\log\ N2^{Mp})$ bits.
- Send one bit to identify the sign of $c_j(f)$ for each $j \in \Lambda_k$, $0 \leq k \leq M$. This will require $O(2^{Mp})$ bits.
- Send $m$ bits to describe each $c_j(f)$,$j \in \Lambda_k$, for $0 \leq k \leq M$. This will require $O(m2^{Mp})$ bits.

Thus the total number of bits used in the encoding is $O(m2^{Mp})$.

Notice that for each $j \in \Lambda_k$, $0 \leq k \leq M$, we can recover each $c_j(f)$ by

$$c_j = \pm \sum_{i=0}^{m} b_i 2^{-k-i}$$

where the sign is given by the sign bit. It follows that
$\mid c_j(f) - c_j \mid \leq 2^{-m-k}$ for every such coefficient. Here we have used the

fact that knowing that $j \in \Lambda_k$ means that the first nonzero binary bit of $c_j(f)$ is the $k$-th bit.

To decode we simply set

$$f = \sum_{k=0}^{M} \sum_{j \in \Lambda_k} c_j \psi_j$$

We now analyze the error we have incurred in such an encoding. The square of the error will consist of two parts. The first corresponds to the $j \in \Lambda_k$, $0 \leq k \leq M$. For each such $j$ we have $\mid c_j(f) - c_j \mid \leq 2^{-m-k}$ and so the total square error for this is

$$\leq C \sum_{k=1}^{M} 2^{kp} 2^{-2m} 2^{-2k} \leq c 2^{-2m}$$

because $p \leq 2$. The second part of the error corresponds to all the coefficients which have magnitude $\leq 2^{-M}$. We have that this sum does not exceed $\sum_{|c_j| > 2^{-M}} \mid c_j \mid^2 \leq 2^{-M(2-p)} \sum_{j=1}^{\infty} \mid c_j \mid^p \leq 2^{-2m}$. Thus the total error we incur is $O(2^{-m})$.

In summary, by allocating $O(m 2^{\frac{m}{1/p-1/2}})$ bits we achieve distortion $C 2^{-m}$. Equivalently, by allocating $n \ \log \ n$ bits, we achieve distortion $C n^{-(1/p-1/2)}$.

**Remark**

This is within a logarithmic factor of the optimal encoding given by Kolmogorov entropy of the class $U(X_p) \cap Y^{\alpha}$. A slightly more careful argument can remove this logarithm.

**Example:**
**The Wavelet Basis**

In the method above we failed to achieve the optimal performance because of the cost involved in identifying which indices were in each $\Lambda_k$. We will now describe a method that can do better, using the Haar basis for $L_2[0,1]$. Thus, we first define the scaling function

$$\varphi := \chi_{[0,1]}.$$

Next, we define the mother wavelet

$$\psi := \chi_{[0,\frac{1}{2}]} - \chi_{[\frac{1}{2},1]}.$$

We then define the remaining wavelets recursively. They are obtained by dilations and shifts of the mother wavelet on dyadic intervals:

$$\psi_J := \frac{2^k}{2}\psi_{[0,1]}(2^k x - j)$$

where $J = [j2^{-k},(j+1)2^{-k}]$ are dyadic intervals. We denote by $D_+$ the collection of all dyadic intervals contained in $[0,1]$. Then, the collection of functions $\{\varphi\} \cup \{\psi_J\}_{J \in D^+}$ forms an orthonormal basis for $L_2[0,1]$.
A key property of wavelets is that a tree structure can be placed on the coefficients due to the use of dyadic intervals in their construction. Thus, let

$$T_k := \{j : |\ c_j\ | \geq 2^{-k}\}$$

and

$$T_{k+1} - T_k = \Lambda_k.$$

We define $T_k$ as the smallest tree containing $T_k$. Given any binary tree of size $n$, we can encode the tree with at most $O(n)$ bits, in the process outperforming the encoder described in above.

Sparse Approximation and ℓp Spaces

We now look at how well $f \in \mathbf{X}$ can be approximated by $n$ functions in the dictionary $\mathscr{D}$.

**Definition**

We define the error of $n$-term approximation of $f$ by the elements of the dictionary $\mathscr{D}$ as

$$(1) \quad \sigma_n(f)_{\mathbf{X}} := \sigma_n(f, \mathscr{D})_X := \inf_{s \in \Sigma_n} \| f - s \|_{\mathbf{X}}.$$

We also define the class of $r$-smooth signals in $\mathscr{D}$ as

$$(2) \quad \mathscr{A}^r := \mathscr{A}^r(\mathscr{D}) := \{f \in X, \sigma_n(f) \leq Mn^{-r} \text{ for some } M\}$$

,

with the corresponding norm $\| f \|_{\mathscr{A}^r} = \sup_{n=1,2,\dots} n^r \sigma_n(f)_{\mathbf{X}}$.

In general, the larger $r$ is, the 'smoother' the function $s \in \mathscr{A}^r(\mathscr{D})_{\mathbf{X}}$. Note also that $\mathscr{A}^r \subseteq \mathscr{A}^{r'}$ if $r > r'$. Given $f$, let $r(f) = \sup\{r : f \in \mathscr{A}^r\}$ be a measure of the "smoothness" of $f$, i.e. a quantification of compressibility.

Let $\mathbf{X} = H$, a Hilbert space[footnote] such as $\mathbf{X} = L_2(\mathbb{R})$, and assume $\mathscr{D} = B$ - an orthonormal basis on $\mathbf{X}$; i.e. if $B = \{\phi_i\}_i$, then $\langle \phi_i, \phi_j \rangle = \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta. This also means that each $f \in \mathbf{X}$ has an expansion $f = \sum_j c_j(f)\phi_j$, where $c_j(f) = \langle f, \phi_j \rangle$. We also have $\| f \|_{\mathbf{X}}^2 = \sum_{j=1}^{\infty} |c_j(f)|^2$.
A Hilbert space is a complete inner product space with the norm induced by the inner product

Recall the definition of $\ell_p$ spaces: let $(a_j) \in \mathbb{R}$; then $(a_j) \in \ell_p$ if $\| (a_j) \|_{\ell_p} < \infty$ with $\| (a_j) \|_{\ell_p} = \left( \sum_j |a_j|^p \right)^{1/p}$ for $p < \infty$ and $\| (a_j) \|_{\ell_p} = \sup_j |a_j|$ for $p = \infty$. We also recall that for $L_p$ spaces on compact sets, $L_p \subset L_{p'}$ if $p > p'$. The opposite is true for $\ell_p$ spaces: $\ell_p \subset \ell_{p'}$ if $p < p'$. Hence, the smaller the value of $p$ is, the "smaller" $\ell_p$ is.

**Example:**

Does there exist a sequence $(a_j)$ with $\| (a_j) \|_{\ell_1} = \sum_j \left|a_j\right| < \infty$ but with $\| (a_j) \|_{\ell_p} = \left(\sum_j |a_j|^p\right)^{\frac{1}{p}} = \infty$ for all $0 < p < 1$? Consider the sequence $a_n = \frac{1}{n(\log n)^{1+\delta}}$. We see that $(a_n) \in \ell_1$ but $\| (a_n) \|_{\ell_p} = \infty$ for all $0 < p < 1$.

A sequence $(a_n)$ is in $\ell_p$ if the sorted magnitudes of the $a_n$ decay faster than $n^{-\frac{1}{p}}$.

Define $a_n^*$ as the element of the sequence $(a_n)$ with the $n^{\text{th}}$ largest magnitude, and denote $\left(a_n^*\right)$ as the decreasing rearrangement of $(a_n)$. It is easy to show that $k\left(a_k^*\right)^p \leq \sum_n (a_n)^p$ for all $k$; also, if $(a_n) \in \ell_p$, then $a_k^* \leq \| (a_n) \|_{\ell_p} k^{-\frac{1}{p}}$.

**Definition**

A sequence $(a_n)$ is in weak $\ell_p$, denoted $(a_n) \in w\ell_p$, if $a_k^* \leq Mk^{-\frac{1}{p}}$. We also define the quasinorm [footnote] $\| (a_n) \|_{w\ell_p}$ as the smallest $M > 0$ such that $a_k^* \leq Mk^{-\frac{1}{p}}$ for each $k$.
A quasinorm is has the properties of a norm except that the triangle inequality is replaced by the condition $\| x + y \| \leq C_0 [\| x \| + \| y \|]$ for some absolute constant $C_0$.

**Example:**
The sequence $a_n = \frac{1}{n}$ is in weak $\ell_1$ but not in $\ell_1$.

For $p$,$p'$ such that $p' > p$, we have $\ell_p \subset w\ell_p \subset \ell_{p'}$.

Let $\mathscr{D} = B$ be an orthonormal basis for the Hilbert space $\mathbf{X} = H$. For $f \in \mathbf{X}$ with representation in $B = [\phi_1, \phi_2, \ldots]$ as $f = \sum_n c_n(f)\phi_n$, we have $f \in \mathrm{A}^r(B)_{\mathbf{X}}$ if and only if the sequence $(c_n(f)) \in w\ell_\tau$, with $\frac{1}{\tau} = r + \frac{1}{2}$. Moreover, there exist $C_0, C_0' \in \mathbb{R}$ such that $C_0' \| (c_n(f)) \|_{w\ell_\tau} \leq \| f \|_{\mathrm{A}^r} \leq C_0 \| (c_n(f)) \|_{w\ell_\tau}$.

**Example:**
Let $r = \frac{1}{2}$. $f \in \mathscr{A}^{\frac{1}{2}}$ if and only if $(c_n(f)) \in w\ell_\tau$, i.e. if $c_n^*(f) \leq Mn^{-1} = \frac{M}{n}$.

We prove the converse statement; the forward statement proof is left to the reader. We would like to show that if $(c_n(f)) \in w\ell_\tau$, then $f \in \mathscr{A}^r$, with $r = \frac{1}{\tau} - \frac{1}{2}$. The best $n$-term approximation of $f$ in $B$ is of the form $s = \sum_{k\in\Lambda} a_k\phi_k, \sharp\Lambda \leq n$. Therefore, we have:

$$\begin{aligned}
\sigma_n(f)_{\mathbf{X}} &= \inf_{s\in\Sigma_n} \| f - s \|_{\mathbf{X}} = \inf_{s\in\Sigma_n} \| \sum_{k\in\Lambda} (c_k(f) - a_k)\phi_k + \sum_{k\notin\Lambda} c_k(f)\phi_k \|_X \\
&= \inf_{s\in\Lambda} \sum_{k\in\Lambda} (c_k(f) - a_k)^2 + \sum_{k\notin\Lambda} (c_k(f))^2 = \sum_{k=n+1}^{\infty} \left|c_k^*(f)\right|^2 \qquad (3) \\
&\leq M^2 \sum_{k=n+1}^{\infty} k^{-\frac{2}{\tau}} \leq M^2 \sum_{k=n+1}^{\infty} k^{-2r-1} \text{ (since } (c_n(f)) \in w\ell_p),
\end{aligned}$$

where $M := \| \left(c_n(f)\right) \|_{w\ell_p}$.

We prove the converse statement; the forward statement proof is left to the reader. We would like to show that if $(c_n(f)) \in w\ell_\tau$, then $f \in \mathscr{A}^r$, with $r = \frac{1}{\tau} - \frac{1}{2}$. The best $n$-term approximation of $f$ in $B$ is of the form $s = \sum_{k\in\Lambda} a_k\phi_k, \sharp\Lambda \leq n$. Therefore, we have:

$$\begin{aligned}
\sigma_n(f)_{\mathbf{X}} &= \inf_{s\in\Sigma_n} \| f - s \|_{\mathbf{X}} = \inf_{s\in\Sigma_n} \| \sum_{k\in\Lambda} (c_k(f) - a_k)\phi_k + \sum_{k\notin\Lambda} c_k(f)\phi_k \|_X \\
&= \inf_{s\in\Lambda} \sum_{k\in\Lambda} (c_k(f) - a_k)^2 + \sum_{k\notin\Lambda} (c_k(f))^2 = \sum_{k=n+1}^{\infty} \left|c_k^*(f)\right|^2 \qquad (3) \\
&\leq M^2 \sum_{k=n+1}^{\infty} k^{-\frac{2}{\tau}} \leq M^2 \sum_{k=n+1}^{\infty} k^{-2r-1} \text{ (since } (c_n(f)) \in w\ell_p),
\end{aligned}$$

where we define $C = \frac{\lambda_0}{r}$. Using this result in the earlier statement, we get

$$\sum_{k=n+1}^{\infty} \left|c_k^*(f)\right|^2 \leq CM^2 n^{-2r} \leq M^2 z n^{-r}; \qquad (5)$$

this implies by definition that $(c_k(f)) \in \mathscr{A}^r$.

Thresholding and Greedy Bases

We shall next discuss some notions related to best $n$-term approximation.

## Thresholding

1. Let $\mathbb{X}$ be a Hilbert space. Given $f$, let $\Lambda_\epsilon(f) = \{j : |c_j(f)| > \epsilon\}$. The thresholding operator $T_\epsilon$ is defined by
   **Equation:**

$$T_\epsilon f = \sum_{j \in \Lambda_\epsilon(f)} C_j(f)\varphi_j.$$

   It is easy to see that for each $\epsilon$, $T_\epsilon f$ is the best approximation to $f$ using $N$ terms where $N$ is the cardinality of $\Lambda_\epsilon$:
   **Equation:**

$$\| f - T_\epsilon \|_{\mathbb{X}} = \sigma_N(f)_{\mathbb{X}}.$$

   Thresholding is easily implemented on a computer.
2. The thresholding scheme above can be generalized if $\mathbb{X}$ is not a Hilbert space provided the dictionary has some specific structure. For example, when

   1. The dictionary is the wavelet basis and $\mathbb{X} = L_p$, $1 < p < \infty$.
   2. $\mathbb{X} = l_p$ and the dictionary is the canonical basis $\delta_j = \varphi_j$. ex: $(0, 0, ..., 1, 0)$
   3. For a general Banach space $\mathbb{X}$ and the dictionary $(\varphi_j)$ is a greedy basis.

## Greedy Bases

We briefly describe the notion of greedy basis.
**Definition**

Given $\mathbb{X}$, we say $(\varphi_j)$ is a greedy basis for $\mathbb{X}$ if for each $\epsilon > 0$,
**Equation:**

$$\| f - T_\epsilon f \|_{\mathbb{X}} \leq C(\mathbb{X}) \sigma_N(f)_{\mathbb{X}}$$

where $N$ is the cardinality of $\Lambda_\epsilon$.
**Definition**

A basis $\varphi_j$ is said to be unconditional if
**Equation:**

$$\| \sum \pm c_j \varphi_j \|_{\mathbb{X}} \leq C \| \sum c_j \varphi_j \|_{\mathbb{X}}$$

or equivalently
**Equation:**

$$\| \sum c_j \varphi_j \|_{\mathbb{X}} \leq C \| \sum d_j \varphi_j \|_{\mathbb{X}} \qquad \text{where} \qquad \left| c_j \right| \leq \left| d_j \right|.$$

This is an older concept from functional analysis. In words, this definition says that if the terms $c_j$ are rearranged, the series $\sum c_j \varphi_j$ will still converge. This is not generally true for all bases.
**Definition**

A basis $\varphi_j$ is said to be democratic if
**Equation:**

$$\| \sum_{j \in \Lambda} \varphi_j \| \leq C \| \sum_{j \in \Lambda'} \varphi_j \|,$$

where the cardinality of $\Lambda'$ equals the cardinality of $\Lambda$.
**Remark**

$(\varphi_j)$ greedy $\leftrightarrow$ $(\varphi_j)$ is both unconditional and democratic.

Some examples involving the last two definitions:

- The fourier basis in $L_p$ is not democratic, but is unconditional for $l < p < \infty$.
- The wavelet basis contains both of these properties, and is therefore greedy.

If $\mathbb{X} = L_p$ has $(\varphi_j)$ greedy, $B = \{\varphi_j\}$, $f = \sum_{j=1}^{\infty} c_j(f)\varphi_j$, $c_j(f) = \langle f, \psi_j \rangle$ where $\psi_j$ is a dual basis,
**Equation:**

$$f \in \mathscr{A}^r(B) \leftrightarrow (c_j(f)) \in w_{l_\tau}, \qquad \frac{1}{\tau} = r + \frac{1}{p}$$

and
**Equation:**

$$\| f \|_{\mathscr{A}^r} \approx \| c_j(f) \|_{l_\tau}$$

Let us now consider a specific setting that we shall be concerned with a lot in this course. We shall examine some of the concepts we have introduced in the finite dimensional space of of all sequence (points) in $\mathbb{R}^N$. Recall that we can put many different norms on this space including the $\ell_p$ norms and the weak $\ell_p$ norms.
**Remark**

Given a vector $= (x_1, x_2, ..., x_N) \in \mathbb{R}^N$. The best approximation to $x$ from $\Sigma_n$ in the $\ell_p$ norm is to take the vector in $\Sigma_n$ which shares the $n$ largest values of $x$. Its error of approximation satisfies
**Equation:**

$$\sigma_n(x)_{\ell_p} \leq Cn^{-r} \| x \|_{w_{l_\tau}}, \qquad \frac{1}{\tau} = r + \frac{1}{p}$$

**Remark**

$\| x \|_{w_{l_\tau}} \leq \| x \|_{l_\tau}, \qquad \frac{1}{\tau} = r + \frac{1}{p}.$

**Example**

For $p = 1$ and $r = 3$, $\sigma_n(x)_{\ell_1} \leq Cn^{-3} \parallel x \parallel_{w_{l_\tau}}$ and $\frac{1}{\tau} = 4$. In words, this equation shows what kind of $\tau$ is needed for a given decay rate (or given some $\tau$, what kind of decay rate will be achieved) to approximate with certain ability.

**Example**

Show $\sigma_n(x)_{l_p} \leq Cn^{-r} \parallel x \parallel_{l_\tau}$ holds with $C = 1$.

Proof: Let $\Lambda_n := \{i : |x_i| \text{largest}\}$,

**Equation:**

$$\sigma_n(x\Big)_{l_p}^p = \sum_{i \notin \Lambda_n} \Big|x_i\Big|^p$$

**Equation:**

$$\leq \sum_{i \notin \Lambda_n} \Big|x_i\Big|^{p-\tau} \Big|x_i\Big|^\tau$$

**Equation:**

$$\leq \Big(\parallel x \parallel_{w_{l_\tau}} n^{-\frac{1}{\tau}}\Big)^{p-\tau} \Big(\sum \Big|x_i\Big|^\tau\Big)$$

**Equation:**

$$\leq \parallel x \parallel_{l_\tau}^{p-\tau} \parallel x \parallel_{l_\tau}^\tau n = n^{-rp} \parallel x \parallel_{l_\tau}^p$$

and so

**Equation:**

$$\sigma_n(x)_{l_p}^p \leq n^{-rp} \parallel x \parallel_{l_\tau}^p$$

$\sigma_n(x)_{l_p} \leq n^{-r} \parallel x \parallel_{l_\tau}$

**Remark**

For $\mathbb{X} = L_p$, $\{\varphi_j\}$ a wavelet basis, we can say wavelet coefficients of $f$ are in $l_p$ is equivalent to $f$ is in a certain Besov class (roughly speaking $f$ has $r$ derivatives and $f^{(r)} \in L_\tau$). We refer the reader to [link] for precise formulations of results of this type.

## Greedy Algorithms

We now turn to the questions of generating good approximations for $n$ term approximation from a general dictionary We shall assume that the dictionary $\mathscr{D}$ is complete in the Hilbert space $\mathbb{H}$. This means that every element in $\mathbb{H}$ can be approximated arbitrarily well by linear combinations of the elements of $\mathscr{D}$. Since the dictionary is no longer an orthogonal basis as was considered above, we need to revisit how to find good $n$ term approximations. Because of redundancy within the dictionary, we cannot simply pick the largest coefficients as we saw with a basis. Greedy algorithms are a method to generate good $n$ term approximations.

1. General Greedy Algorithm Given $f$, we want to generate an $n$-term approximation to $f$.
   **Equation:**

   $$f \to s = \sum_{s=1}^{n} c_j g_j$$

   The general steps are as follows:

   1. Initialize: (approximation) $s_0 = 0$, (residual) $r_0 = f$, approximation collection$\Lambda_0 = \emptyset$
   2. Search $\mathscr{D}$ for some $g \in \mathscr{D}$, then add $g$ to the set $\Lambda$.
   3. Use $\{g_1, g_2, ..., g_n\}$ to find new approximation for $s_n$.

   At stage $n$, we have $s_n$, $r_n = f - s_n$, and $\Lambda_n = \{g_1, g_2, ..., g_n\}$. There are many types of greedy algorithms. We describe the three most common in the case $\mathbb{S}$ is a Hilbert space. However, there are anlaogues of these for $L_p$.
2. Pure Greedy Algorithm (PGA) Note: >From $r_n$ choose $g_{n+1} := \text{argmax}|\langle r_n(f), g\rangle|$ (the $g$ that causes the largest inner product).
   **Equation:**

   $$s_{n+1} = s_n + \langle r_n(f), g\rangle g$$

**Equation:**

$$r_{n+1} = f - s_n - \langle f - s_n, g \rangle g = f - s_{n+1}$$

This method is similar to a steepest decent algorithm for decreasing the error.

3. Orthogonal Greedy Algorithm (OGA) >From $r_n$ choose $g_{n+1} := \text{argmax}|\langle r_n\,(f), g\rangle|$ as in the PGA.
**Equation:**

$$V_{n+1} := sp\{g_1, g_2, ..., g_{n+1}\}$$

**Equation:**

$$s_{n+1} := p_{V_{n+1}} f = \sum_{j=1}^{n+1} \alpha_j g_j$$

where $P_V$ denotes the orthogonal projection onto the space $V$. We can find $s_{n+1} = P_{V_{n+1}} f$ by solving the linear system of equations
**Equation:**

$$\left\langle \sum_{j=1}^{n+1} \alpha_j g_j, g_k \right\rangle = \langle f, g_k \rangle.$$

Then, $r_{n+1} = f - s_{n+1}$.

4. Relaxed Greedy Algorithm (RGA) >From $r_n$ choose $g_{n+1}$ in some way (for example, our earlier methods) and then define
**Equation:**

$$s_{n+1}\,(f) = \alpha s_n + \beta g_{n+1}$$

Unlike PGA, here we do not make a full step in the correct direction. For example, one way to proceed is to define
**Equation:**

$$\text{arginf}_{\alpha,\beta,g} \parallel f - \alpha s_n + \beta g_n \parallel =: \alpha^*, \beta^*, g^*$$

This type of greedy algorithm is known to perform the best as compred with the previous two.

## Measuring Performance

Given $\mathbb{X}$, $\mathscr{D}$, it is not practical to minimize $\sigma_n(f)_{\mathbb{X}}$ by searching over all the possibilities. The greedy approximation gives an $n$-term solution with less computation, but does it perform well?

Let
**Equation:**

$$\mathscr{L}^1\left(\mathscr{D}\right) := \left\{f \in \mathbb{X} : \sum c_g g, \sum_{g\in\mathscr{D}} \middle| c_g \mid \leq M\right\}$$

where the smallest $M$ is the $\mathscr{L}^1$ norm of $f$.

For OGA or RGA as described above, we have
**Equation:**

$$\parallel f - s_n f \parallel \leq C n^{-\frac{1}{2}} |f|_{\mathscr{L}^1}.$$

**Remark**

Remark 5 This is similar to $\sigma_n(x)_{l_2} \leq n^{-\frac{1}{2}} \parallel x \parallel_{l_1}$ ($n$-term approximation) but its not always quite as good.

## Compressive Sensing

We now consider a different setting. Suppose $x \in \mathbb{R}^N$ and we wish to sample $x$, where taking a sample means the application of a linear functional $\lambda \in \mathbb{R}^N$ to $x$. Next, we prescribe a budget of $n$ samples, and consider all linear encoders using $n$ samples. We can write these $n$ linear functionals as an $n \times N$ matrix $\Phi$:$\mathbb{R}^N \to \mathbb{R}^n$. We then consider a decoder $\Delta$:$\mathbb{R}^n \to \mathbb{R}^N$. Our approximation to $x$ is thus $\Delta(\Phi(x))$.

To make the problem precise, we first pick a measure for distortion:

$$\text{error} = \|x - \Delta\left(\Phi\left(x\right)\right)\|_{\ell_p}.$$

We next must make some assumption about $x$. For example, we can assume that

$$x \in \Sigma_k = \{x : x_i = 0 \ \text{ for } \ i \notin \Lambda, \sharp\Lambda \leq k \},$$

or

$$x \in \ell_\tau$$

or

$$x \in w_{\ell_\tau}$$

We recall our basic problem: a signal $x \in \mathbb{R}^N$ will be “sampled” or “sensed” by applying the $n$ linear projections represented by the columns of the sampling matrix $\Phi_{n \times N}$. The resulting measurements are given in the vector $y \in \mathbb{R}^n$, where $y = \Phi x$. We will assume that $n < N$, meaning that in addition to thinking of of the sampling operation $\Phi$ as an encoder, we can also view it as a projection to a lower dimensional linear subspace. In either case, we would like the measurements $y$ to preserve as much information about the signal $x$ as possible. To proceed in finding optimal solutions, we must formalize this problem and define how we will measure this information loss.

Moving forward, a critical quantity for us will be the null space of the sampling matrix, $N = N(\Phi) = \{x : \Phi x = 0\}$. Because we are trying to take as few measurements as possible, we will assume that we are taking measurements efficiently so that the rows of $\Phi$ are all linearly independent. Therefore, assume that $\text{rank}(\Phi) = n$, implying that $N$ has dimension $(N - n)$. The non-trivial null space of $\Phi$ means that it is not an invertible mapping. For any measurement vector $y$ we can define the class of all observable signals that would result in the same measurement, $\mathscr{F}(y) := \{x : \Phi x = y\} \subseteq \mathbb{R}^N$. The class $\mathscr{F}(y)$ can always be written as a sum of a vector in the class and a vector in the null space of the sampling matrix, $\mathscr{F}(y) = x_0 + N$ , where $x_0 \in \mathscr{F}(y)$. If we have two vectors $x_0$,$x_1 \in \mathscr{F}(y)$, then by linearity we know that $\Phi(x_1 - x_0) = 0$. This fact implies that $(x_1 - x_0) \in N$, and consequently that $x_1 \in x_0 + N$.

Associated to our encoder $\Phi$, we shall have to describe a decoder $\Delta$, which is a (not necessarily linear) map $\Delta : \mathbb{R}^n \to \mathbb{R}^N$. This decoder will take the measurements $y$ and try to recover $x$ as closely as possible, $x = \Delta(\Phi x)$. In order to design the best decoder $\Delta$, we must specify our metric for measuring the quality of the estimate $x$. For the moment we shall always think of taking an optimal decoder for the problem at hand. Later we shall discuss specific and concrete decoders.

To begin our discussion of the efficiency of the enoder $\Phi$ we consider the following problem.

**Note:** We will fix $n$ and $N$ and try to find $\Phi$ and a decoder $\Delta$ such that for all input signals in a sparsity class $x \in \Sigma_k$ we can get perfect reconstruction, $\Delta(\Phi x) = x$. We will be interested in determining what the largest value of $k$ is that we can find such an encoder/decoder pair.

An important role in this problem and later problems of compressed sensing is played by certain submatrices of $\Phi$. Given a set $T \subseteq \{1,2,\ldots,N\}$, representing a collection of column indices we define the matrix $\Phi_T$ as the

one formed from $\Phi$ by using the columns from the set $T$. The matrix $\Phi_T$ is a $(n \times \#(T))$ matrix. But sometimes we will also use the same notation $\Phi_T$ to denote the matrix obtained from $\Phi$ by setting all entries not in the columns of $T$ to zero.

We can now state the following theorem.

The following statements are all equivalent for any given $\Phi$:

1. There exists a decoder $\Delta$ such that $\Delta(\Phi x) = x$ for all $x \in \Sigma_k$.
2. $N(\Phi) \cap \Sigma_{2k} = \{0\}$.
3. $\Phi_T$ has rank $\#(T)$ for all $T$ with $\#(T) = 2k$.
4. ${\Phi_T}^t \Phi_T$ is non-singular (i.e., invertible) for all $T$ with $\#(T) = 2k$.

The equivalence of $b \leftrightarrow c \leftrightarrow d$ is simple linear algebra. First let us prove that $a \rightarrow b$. Assume $a$. Suppose that there was a vector $\eta \in N \cap \Sigma_{2k}$. We know that we can write $\eta = x_0 - x_1$, where $x_0$,$x_1$ both have support less than $k$. We could write $\eta$ simply as a composite vector with support $2k$ where the first half of $\eta$ is $x_0$ and the second half of $\eta$ is $x_1$, $\eta = (x_0 \mid x_1)$. We know that $\eta \in N$, which implies that $\Phi\eta = 0$. It then follows that $\Phi x_0 = \Phi x_1$, and we have as a consequence that $\Delta\Phi x_0 = \Delta\Phi x_1$ and finally that $x_0 = x_1$. This proves that $\eta = 0$ as desired.

Now let us prove that $b \rightarrow a$. Suppose that we have a measurement $\Phi x = y \in \mathbb{R}^n$, where $x \in \Sigma_k$. We will define the decoder $\Delta(y)$ to be the signal $x \in \mathscr{F}(y)$ with smallest support. Since $x$ has support $k$ so will $x$. We claim that there is no other $x' \in \mathscr{F}(y) \cap \Sigma_k$. Indeed, if $x'$ existed, then $x - x' \in N \cap \Sigma_{2k}$. But, the only vector in $N \cap \Sigma_{2k}$ is zero, implying that $x = x'$. This finally gives us that $\Delta(\Phi x) = x$. $\square$

Using the previous theorem, we can turn to the question of finding good encoder/decoder pairs.[footnote] Given a fixed $N$, how large can $k$ be, and what is the best $\Phi$? Given a fixed $n$, the largest $k$ is $k = \lfloor \frac{n}{2} \rfloor$. Alternatively, we can say that given a fixed $k$, we need at least $n = 2k$ measurements. Another way to say this is that there exist encoding matrices $\Phi_{2k \times N}$ such that any selection of $2k$ columns are linearly independent. Examples are the

DFT matrix or the Vandermonde matrix corresponding to interpolation at distinct points $z_1, \ldots, z_N$.

A question about whether we will ever really find natural signals in $\Sigma_k$ brings to ming a story... Once upon a time, a man was floating over the countryside in a hot air balloon. The man in the balloon yelled down to a stranger on the ground and asked “Where am I?” The man on the ground thought for about 5 minutes and then answered “You’re in a hot air balloon.” The man in the balloon responded with “You must be a mathematician,” to which the man on the ground answered “Yes, how did you know?” “Because,” replied the man in the balloon, “you had to think a long time before you answered, your answer was very precise, and your answer was completely useless!” So, yes, we may be dealing with a limited model, but we have to crawl before we can walk.

Gelfand n-widths

Continuing from last time, we have a signal $x \in \mathbb{R}^N$, an $n \times N$ measurement matrix $\Phi$, and $y = \Phi x \in \mathbb{R}^n$ is the information we draw from $x$.

We consider decoders $\Delta$ mapping $\mathbb{R}^n \to \mathbb{R}^N$. We have been discussing whether there exists a decoder with certain properties. So for this discussion (about information preservation), we can just think about optimal decoding.

While the previous result on sparse signal recovery is interesting, it is not very robust. What happens if our signal does not have support $k$? Can we still obtain meaningful results? The answer is yes. To do so we introduce more general input signal classes $K \subseteq X$ that allows fully supported signals. For example, we will consider the signal class defined by the unit ball
**Equation:**

$$K = U\left(\ell_p^N\right) = \left\{x : \| x \|_X \leq 1\right\}.$$

Given an encoder/decoder pair $(\Phi, \Delta)$, the worst case error on a set $K$ for that pair will be given by
**Equation:**

$$E(K,\Phi,\Delta)_X = \sup_{x \in K} \| x - \Delta(\Phi x) \|_X.$$

Finally, using min-max principles we will define the minimum error over all encoder/decoder pairs for a signal class and for a fixed number of measurements $n$ to be
**Equation:**

$$E_n(K)_X = \inf_{(\Phi,\Delta):\Phi \text{ is } n \times N} E(K,\Phi,\Delta)_X.$$

This measure $E_n(K)_X$ is the best we could do while measuring distortion on the topology of $X$, using $n$ linear measurements, and using arbitrary decoding.

We will see that these questions are actually equivalent to a classical "$n$-width" problem. $n$-widths have seen a great deal of work over the years by a variety of mathematicians: Kolmogorov, Tikhomirov, Kashin, Gluskin, etc. There are many different flavors of $n$-widths, but we will study the Gelfand $n$-width (the least intuitive of the widths).

**Definition**

Gelfand n-width

Let $K \subseteq X$ be compact. Given $n$, the Gelfand width (also called the dual width) is given by

**Equation:**

$$d^n(K)_X := \inf_{Y:\,\mathrm{codim}(Y)=n} \sup \left\{ \| x \|_X : x \in K \cap Y \right\}.$$

where by codimension $(Y)=n$ we mean that $Y$ has dimension $\dim(X) - n$.

In other words, we are looking for the subspace $Y$ that slices through the set $K$ so that the norms of the projected signals are as small as possible. We can now state the following theorem about n-widths:

Provided that $K$ has the properties (1) $K = -K$ and (2) $K + K = CK$, then

**Equation:**

$$d^n(K)_X \leq E_n(K)_X \leq C d^n(K)_X$$

where $C$ is the same constant listed in property (2).[footnote]

Clarifying notation: $CK = \{Cx : x \in K\}$ and $K + K = \{x_1 + x_2 : x_1, x_2 \in K\}$.

We start with the left-hand inequality. We want to take any encoder/decoder pair and use that to construct a $Y$. So let $\Phi$, $\Delta$ be an encoder/decoder. Then simply let $Y = \mathcal{N}(\Phi)$. Now consider an $\eta \in K \cap Y$ and note that $\Phi(\eta) = 0$ since $\eta \in Y$. Let $z = \Delta(0)$ be the decoding of 0 (practically speaking, $z$ should be zero itself, but we avoid that assumption in this proof). Then
**Equation:**

$$\begin{aligned} \| \eta \|_X &\leq \max\left(\| \eta - z \|_X, \| \eta + z \|_X\right) \\ &= \max\left(\| \eta - \Delta\Phi(\eta) \|_X, \| -\eta - \Delta\Phi(\eta) \|_X\right) \\ &\leq \sup_{x \in K} \| x - \Delta\Phi(x) \|_X \end{aligned}$$

where we first employ the triangle inequality, then the fact that multiplying by $-1$ does not change the norm, then the fact that $K = -K$. So then
**Equation:**

$$\sup_{\eta \in K \cap Y} \| \eta \|_X \leq \sup_{x \in K} \| x - \Delta\Phi(x) \|_X.$$

Taking the infimum over all $\Phi$, $\Delta$, it follows that
**Equation:**

$$d^n(K)_X \leq E_n(K)_X.$$

Since $\Phi$ is $n \times N$, then $\dim(\mathcal{N}(\Phi)) \geq N - n$.

Now we prove the right-hand inequality. Assume we have a good $Y$. Suppose $Y$ has codimension $N - n$. Then $Y^\perp$ (the orthogonal complement of $Y$ in $\mathbb{R}^N$) has dimension $n$. Let $v_1, v_2, \cdots, v_n \in \mathbb{R}^N$ be a basis for $Y^\perp$. Let $\Phi$ be the $n \times N$ matrix obtained by stacking the rows $v_1, v_2, \cdots, v_n$. Then $\mathcal{N}(\Phi) = Y$. Define $\Delta(y) =$ any element of $K \cap \mathcal{F}(y)$ if there is one (otherwise let $\Delta(y)$ be anything in $\mathcal{F}(y)$). Now look at the performance $\| x - \Delta\Phi(x) \|_X$ for some $x \in K$. Both $x$ and $\Delta\Phi\left(x\right) =: x'$

are elements of $K$, so $x - x'$ is in $\mathcal{N}(\Phi)$ and in $CK$. Therefore $\frac{x-x'}{C} \in K \cap \mathcal{N}(\Phi)$. Thus,
**Equation:**

$$\| \frac{x - x'}{C} \|_X \leq \sup_{z \in Y \cap K} \| z \|_X,$$

and so for any $x \in K$,
**Equation:**

$$\| x - \Delta\Phi\left( x \right) \|_X \leq C \sup_{z \in Y \cap K} \| z \|_X.$$

Taking the infimum over all $Y$, we get that $E_n(K)_X \leq C d^n(K)_X$.

From the proof of this theorem, we see that there is a matching between the matrices $\Phi$ and the spaces $Y$ (via the nullspace of $\Phi$).

An important result is that $d^n\left(U\left(\ell_q^N\right)\right)_{\ell_p^N}$ is known for all $p, q$ except $p = 1, q = \infty$. A precise statement of these widths can be found in the book [link]. A particularly important case is
**Equation:**

$$C_1 \sqrt{\frac{\log(N/n)}{n}} \leq d^n\left(U\left(\ell_1^N\right)\right)_{\ell_2^N} \leq C_2 \sqrt{\frac{\log(N/n)}{n}}$$

for $N > 2n$. This result was first proved by Kashin with a worse logarithm in the upper inequality and later brought to the present form by Gluskin. This result solves several important problems in functional analysis and approximation.

## Instance Optimality

Now we consider another way (actually two related ways) to measure optimality of an encoder/decoder pair.

1. Instance optimality. Suppose we are in $\mathbb{R}^N$ with an $n \times N$ measurement matrix $\Phi$ and a decoder $\Delta$ . Recall that
   **Equation:**

   $$\sigma_k(x)_X := \inf_{z:\#z\leq k} \| x - z \|_X .$$

   We say that the encoding/decoding strategy $\Phi,\Delta$ is instance optimal of order $k$ with constant $C_0$ if
   **Equation:**

   $$\| x - \Delta\Phi(x) \|_X \leq C_0 \sigma_k(x)_X$$

   for all $x \in \mathbb{R}^N$ . (Note that we are no longer restricting $x$ to a class $K$ .) Better $\Phi$ 's have larger $k$ for which this holds. The name "instance optimal" indicates that the encoding/decoding performance depends on each instance of $x$ .
2. Mixed-norm instance optimality (MNIO). Let $q < p$ . The encoder/decoder pair $\Phi,\Delta$ is MNIO for $p,q,k$ , and $C_0$ if
   **Equation:**

   $$\| x - \Delta\Phi(x) \|_{\ell_p^N} \leq C_0 \frac{\sigma_k(x)_{\ell_q^N}}{k^{1/q-1/p}}.$$

   Cases of interest include asking whether
   **Equation:**

   $$\| x - \Delta\Phi(x) \|_{\ell_2^N} \leq C_0 \sigma_k(x)_{\ell_1^N}.$$

   and whether
   **Equation:**

   $$\| x - \Delta\Phi(x) \|_{\ell_2^N} \leq C_0 \frac{\sigma_k(x)_{\ell_1^N}}{\sqrt{k}}.$$

Let’s focus on instance optimality. It would be interesting to know whether a given $\Phi$ satisfies this property. To answer this question, we state an equivalent condition to instance optimality.

Consider the statements

1. $\Phi,\Delta$ is instance optimal of order $k$ on $X$ .
2. $\Phi$ has the following nullspace property (NSP):
   $\| \eta \|_X \leq C_1 \| \eta_{T^c} \|_X \; \forall \eta \in N(\Phi), \# T \leq k.$
3. $\| \eta \|_X \leq C_1 \sigma_k(\eta)_X \forall \eta \in N(\Phi).$
4. $\| \eta_T \|_X \leq C_1' \| \eta_T^c \|_X \; \forall \eta \in N(\Phi), \# T \leq k.$

Then (b) and (c) are equivalent with the same constant; (d) is equvalent to (b) and (c) but with a different constant. Also (a) with a value $k$ implies (b) with the same $k$ , and (b) with a value $2k$ implies (a) with a value $k$ .

## The Restricted Isometry Property

We say that an $n \times N$ matrix $\Phi$ has the restricted isometry property (RIP) for $k$ if for each $T \subseteq \{1, \ldots, N\}$ such that $\#T \leq k$ , $\Phi_T$ (the matrix formed by choosing the columns of $\Phi$ whose indices are in $T$ ) has the property

| | |
|---|---|
| $(1-\delta_k)\| x_T \|_{\ell_2} \leq \| \Phi_T(x) \|_{\ell_2} \leq (1+\delta_k)\| x_T \|_{\ell_2}$ | (RIP) |

where $0 < \delta_k < 1$ . This useful definition is by Candes and Tao. The idea is that the embedding of a $k$ -dimensional space in $M$ -dimensional space almost preserves norm – like an isometry. Another way of looking at it is to consider the matrix $\Phi_T^t \Phi_T$ , of size $k \times k$ . This matrix is symmetric, positive definite, and it's eigen-values are between $1-\delta_k$ and $1+\delta_k$ .

I prefer the following modified condition (dubbed the MIRP), which is more convenient for mathematical analysis:

| | |
|---|---|
| $(c_1)^{-1}\| x_T \|_{\ell_2} \leq \| \Phi_T(x) \|_{\ell_2} \leq c_1\| x_T \|_{\ell_2}$ | (MRIP) |

We can now state the following theorem.

If $\Phi$ satisfies MRIP for $2k$ then $\exists \Delta$ s.t. $(\Phi, \Delta)$ is instance optimal for $\ell_1^N$ for $K$ .

This shows that whenever we have a matrix $\Phi$ satisfying the MRIP for $2k$ then it will perform well on encoding vectors (at least in the sense of $\ell_1^N$ accuracy). The question is how can we construct measurement matrices with this property? We can construct $\Phi$ using Gaussian entries and then normalizing the columns.

$\exists$ constant $c > 0$ s.t. if $k \leq c\frac{n}{\log(N/n)}$ then with high probability $\Phi$ satisfies RIP and MRIP for $k$ .

Given $N$ and $n$ , the range of $k$ in the above results reflects how accurately we can recover data. There is another constant $c'$ that serves as a converse bound for Theorem 3. This converse can be derived using Gluskin widths.

**Remark**

The following generic problem is of great interest: Consider the class of matrices $= \{\Phi M \times N, \Phi \text{has some prescribed property}(\text{eg. Toeplitz, circulant, etc.})\}$ . What is the largest $k$ for which such a matrix can have the MRIP.

The Nullspace Property

We begin with a property of the null space $N$ which is at the heart of proving results on instance-optimality.

We say that $N$ has the Null Space Property if for all $\eta \in N$ and all $T$ with $\#T \leq k$ we have $\| \eta \|_X \leq c_1 \| \eta_{T^C} \|_X$

Intuitively, NSP implies that for any vector in the nullspace the energy will not be concentrated in a small number of entries.

The following are equivalent formulations for NSP $X$ for $k$ :

1. $\| \eta \|_X \leq c_1 \sigma_k(\eta)$
2. $\| \eta_T \|_X \leq c_1' \| \eta_{T^C} \|_X$ where $\eta = \eta_T + \eta_{T^C}$ .

Note also that the triangle inequality can be used as follows

$$\| \eta \|_X = \| \eta_T + \eta_{T^C} \|_X \leq \| \eta_T \|_X + \| \eta_{T^C} \|_X$$

which shows that (b) is equivalent to NSP.

1. If $(\Phi, \Delta)$ is instance optimal on $X$ for the value $k$ , then $\Phi$ satisfies the NSP for $2k$ on $X$ with an equivalent constant.
2. If $\Phi$ has the NSP for $X$ and $2k$ then $\exists \Delta$ s.t. $\Phi$ has the instance optimal property for $k$ .

We will prove a slightly weaker version of this to save time. We first prove that instance optimality for $k$ implies NSP $X$ for $k$ (hence this is slightly weaker than advertised) . Let $\eta \in N$ and set $z = \Delta(0)$ then
**Equation:**

$$\begin{aligned} \| \eta - z \| &\leq c_0 \sigma_k(\eta) && \text{instance optimal property} \\ \| \eta + z \| &\leq c_0 \sigma_k(\eta) && -z \in \mathcal{N} \\ \| \eta \| \leq \max\{\| \eta - z \|, \| \eta + z \|\} &\leq c_0 \sigma_k(\eta) && \text{triangle inequality} \end{aligned}$$

We now prove 2. Suppose $\Phi$ has the NSP for $2k$ . Given $y$ , $\mathcal{F}$ $(y) = \{x : \Phi(x) = y\}$. Let us define the decoder $\Delta$ by

$\Delta(y) := \arg\min\{\sigma_K(x)_X : x \in \mathcal{F}(y)\}$ , then

$$
\begin{aligned}
&\| x - \Delta(\Phi(x)) \|_X = \| x - x' \|_X \\
&\leq c_1 \sigma_{2K}(x - x') \\
&\leq c_1(\sigma_K(x) - \sigma_K(x')) \qquad \text{specific 2K term approximation} \\
&\leq 2c_1 \sigma_K(x)
\end{aligned}
$$

QED.

Note that the instance optimal property automatically gives reproduction of $K$ -sparse signals.

At this stage the challenge is to create $\Phi$ with this instance optimal property. For this we shall use the restricted isometry property as introduced earlier and which we now recall.

Optimality and the MRIP

From our last lecture, we are interested in signals $x \in \mathbb{R}^N$, and we can make $n$ measurements (or ask $n$ questions) to obtain $y = \Phi x \in \mathbb{R}^n$. We proposed several optimality criteria to make these measurements, i.e., to choose the measurement matrix $\Phi$.

1. For signals $x \in \Sigma_k$ (with $k$ non-zero coefficients), we choose an encoder $\Phi$ and the corresponding decoder, $\Delta$, such that $\Delta(\Phi(x)) = x$.
2. For classes of signals e.g. $K = U\left(\ell_p^N\right)$, our performance measure is closely related to the Gelfand widths.
3. Instance Optimal: our encoder and decoder, $\Phi$ and $\Delta$, should satisfy

**Equation:**

$$||x - \Delta\Phi(x)||_{\ell_p} \leq C_0 \sigma_k(x)_{\ell_p}.$$

We see that criterion (3) implies (1) since $\sigma_k(x)_{\ell_p} = 0$ for $x \in \Sigma_k$. We also showed in the previous lecture that we have instance optimality for order $k$ if and only if we have the Null Space Property (NSP) of order $2k$. As a reminder, NSP of order $2k$ means that $\forall \eta \in \mathcal{N}, \mathcal{N} = x : \Phi(x) = 0$, we have $\| \eta \|_{\ell_p} \leq C_1 \sigma_{2k}(\eta)_{\ell_p}$. In other words, elements of $\eta \in \mathcal{N}$ are not sparse, and they are all of approximately equal size (i.e., they do not concentrate their entries in $2k$ positions).

We mentioned that, in order to attain instant optimality, the $n \times N$ measurement matrix $\Phi$ should have the modified restricted isometry property (MRIP) of order $m$, i.e., when we choose any $m$ columns of $\Phi$ to obtain a $n \times m$ sub-matrix $\Phi_T$ where $m \leq n$,
**Equation:**

$$C_2^{-1} ||x_T||_{\ell_2} \leq ||\Phi_T x_T||_{\ell_2} \leq C_2 ||x_T||_{\ell_2}.$$

If $\Phi$ has MRIP for $m = 3k$, then $\Phi$ has NSP for $\ell_1$ of order $2k$, and so $\Phi$ is instance optimal in $\ell_1$ of order $k$, i.e.,

**Equation:**

$$||x - \Delta\Phi(x)||_{\ell_1} \leq C\sigma_k(x)_{\ell_1}.$$

A related result is that under the same assumption,
**Equation:**

$$||x - \Delta\Phi(x)||_{\ell_2} \leq \frac{C\sigma_k(x)_{\ell_1}}{\sqrt{k}}.$$

To prove ([link]), we need to only show NSP for $\ell_1$ and $2k$, i.e.,
**Equation:**

$$||\eta||_{\ell_1} \leq C\sigma_{2k}(\eta)_{\ell_1}, \eta \in \mathcal{N}.$$

Given $\eta$, let $T_0$ be the set of indices of the $2k$ largest entries, $T_1$ be the set of indices of the $k$ next largest entries, $T_2$ be the set of indices of the $k$ next largest entries, and so on.
**Equation:**

$$\begin{aligned}
\eta_0 &= \eta_{T_0} + \eta_{T_1}, \\
\eta &= \eta_{T_0} + \eta_{T_1} + ... + \eta_{T_s}, \\
\Phi(\eta) &= \Phi(\eta_{T_0} + \eta_{T_1} + ... + \eta_{T_s}) = 0, \\
\Phi(\eta_0) &= -\Phi(\eta_{T_2} + ... + \eta_{T_s}), \text{by linearity} \\
&= -\sum_{j=2}^{s} \Phi(\eta_{T_j}).
\end{aligned}$$

Therefore, we can estimate
**Equation:**

$$
\begin{aligned}
||\eta_0||_{\ell_2} &\leq C_2||\Phi(\eta_0)||_{\ell_2}, \text{by restricted isometry} \\
&\leq C_2 \sum_{j=2}^{s} \left|\left|\Phi\left(\eta_{T_j}\right)\right|\right|_{\ell_2}, \text{by triangle inequality} \\
&\leq C_2^2 \sum_{j=2}^{s} \left|\left|\eta_{T_j}\right|\right|_{\ell_2}, \text{by restricted isometry.}
\end{aligned}
$$

Since, for $j \geq 2$, $\eta_{T_j}$ is the best k-term approximation to $\eta_{T_{j-1}} + \eta_{T_j}$,
**Equation:**

$$
\left|\left|\eta_{T_j}\right|\right|_{\ell_2} = \sigma_k\left(\eta_{T_{j-1}} + \eta_{T_j}\right)_{\ell_2}.
$$

Furthermore, we know for any $q < p$,
**Equation:**

$$
\sigma_k(\eta)_{\ell_p} \leq k^{\frac{1}{p}-\frac{1}{q}}||x||_{\ell_q}.
$$

Combining ([link]) and ([link]), we obtain
**Equation:**

$$
\left|\left|\eta_{T_j}\right|\right|_{\ell_2} \leq \frac{\left|\left|\eta_{T_{j-1}} + \eta_{T_j}\right|\right|_{\ell_1}}{\sqrt{k}}.
$$

Substituting ([link]) back into ([link]), we now have
**Equation:**

$$
\begin{aligned}
||\eta_0||_{\ell_2} &\leq \frac{C_2^2}{\sqrt{k}} \sum_{j=2}^{s} \left|\left|\eta_{T_{j-1}} + \eta_{T_j}\right|\right|_{\ell_1} \\
&\leq \frac{C_2^2}{\sqrt{k}} \sum_{j=2}^{s} \left|\left|\eta_{T_{j-1}}\right|\right|_{\ell_1} + \left|\left|\eta_{T_j}\right|\right|_{\ell_1} \\
&\leq \frac{2C_2^2}{\sqrt{k}} \sum_{j=1}^{s} \left|\left|\eta_{T_j}\right|\right|_{\ell_1} \\
&\leq \frac{2C_2^2}{\sqrt{k}} \sigma_{2k}(\eta)_{\ell_1}.
\end{aligned}
$$

The last step is due to the fact that $\sum_{j=1}^{s} \left|\left|\eta_{T_j}\right|\right|_{\ell_1}$ is the $2k$-term approximation error for $\eta$. Notice this is only true for $\ell_1$. This completes the proof of ([link]).

To prove ([link]), we let $x_j$ be the $j$-th entry in $\eta_{T_0}$. By the Cauchy-Schwarz Inequality,
**Equation:**

$$
\begin{aligned}
||\eta_{T_0}||_{\ell_1} &= \sum_{j=1}^{2k} \left| x_j \right| \\
&\leq \left( \sum_{j=1}^{2k} 1^2 \right)^{\frac{1}{2}} \left( \sum_{j=1}^{2k} \left| x_j \right|^2 \right)^{\frac{1}{2}}, \text{by CSI} \\
&= \sqrt{2k} ||\eta_{T_0}||_{\ell_2} \\
&\leq \sqrt{2k} ||\eta_0||_{\ell_2} \\
&\leq 2\sqrt{2} C_2^2 \sigma_{2k}(\eta)_{\ell_1},
\end{aligned}
$$

The $2k$-term approximation error in $\ell_1$ for $\eta$ can be expressed as
**Equation:**

$$||\eta_{T_1} + ... + \eta_{T_s}||_{\ell_1} = \sigma_{2k}(\eta)_{\ell_1}.$$

Since
**Equation:**

$$\eta = \eta_{T_0} + (\eta_{T_1} + ... + \eta_{T_s}),$$

we can finally prove that
**Equation:**

$$\begin{aligned} ||\eta||_{\ell_1} &\leq ||\eta_{T_0}||_{\ell_1} + ||\eta_{T_1} + ... + \eta_{T_s}||_{\ell_1}, \text{by triangle inequality} \\ &= 2\sqrt{2}C_2^2\sigma_{2k}(\eta)_{\ell_1} + \sigma_{2k}(\eta)_{\ell_1} \\ &= (2\sqrt{2}C_2^2 + 1)\sigma_{2k}(\eta)_{\ell_1}. \end{aligned}$$

Therefore, we have proved ([link]) with $C = 2\sqrt{2}C_2^2 + 1$.$\square$

Summary

## Review

Last time we proved that for each $k \leq c_0 \frac{n}{\log N/n}$, there exists an $n \times N$ matrix $\Phi$ and a decoder $\Delta$ such that

- **(a)** $\| x - \Delta\Phi(x) \|_{\ell_1} \leq c_0 \sigma_k(x)_{\ell_1}$
- **(b)** $\| x - \Delta\Phi\left(x\right) \|_{\ell_2} \leq c_0 \frac{\sigma_k(x)_{\ell_1}}{\sqrt{k}}$

Recall that we can find such a $\Phi$ by setting the entries $[\Phi]_{j,k} = \varphi_{j,k}(\omega)$ to be realizations of independent and identically distributed Gaussian random variables.

## Deficiencies

### Decoding is not implementable

Our decoding "algorithm" is:
**Equation:**

$$\Delta\left(y\right) := \text{argmin}_{x \in \mathscr{F}(y)} \sigma_k(x)_{\ell_1}$$

where $\mathscr{F}(y) := \{x : \Phi(x) = y\}$. In general, this algorithm is not implementable. This deficiency, however, is easily repaired. Specifically, define
**Equation:**

$$\Delta_1\left(y\right) := \text{argmin}_{x \in \mathscr{F}(y)} \| x \|_{\ell_1}.$$

Then (a) and (b) hold for $\Delta_1$ in place of $\Delta$. This decoding algorithm is equivalent to solving a linear programming problem, thus it is tractable and can be solved using techniques such as the interior point method or the simplex method. In general, these algorithms have computational

complexity $O\left(N^3\right)$. For very large signals this can become prohibitive, and hence there has been a considerable amount of research in faster decoders (such as decoding using greedy algorithms).

**We cannot generate such Φ**

The construction of a $\Phi$ from realizations of Gaussian random variables is guaranteed to work with high probability. However, we would like to know, given a particular instance of $\Phi$, do (a) and (b) still hold. Unfortunately, this is impossible to check (since, to show that $\Phi$ satisfies the MRIP for $k$, we need to consider all possible submatrices of $\Phi$). Furthermore, we would like to build $\Phi$ that can be implemented in circuits. We also might want fast decoders $\Delta$ for these $\Phi$. Thus we also may need to be more restrictive in building $\Phi$. Two possible approaches that move in this direction are as follows:

1. Find $\Phi$ that we can build such that we can prove instance optimality in $\ell_1$ for a smaller range of $k$, i.e.,
   **Equation:**

   $$\| x - \Delta\Phi(x) \|_{\ell_1} \leq c_0 \sigma_k(x)_{\ell_1}$$

   for $k < K$. If we are willing to sacrifice and let $K$ be smaller than before, for example, $K \approx \sqrt{n}$, then we might be able to prove that $\Phi_T^t \Phi_T$ is diagonally dominant for all $T$ such that $\sharp T = 2k$, which would ensure that $\Phi$ satisfies the MRIP.
2. Consider $\Phi(\omega)$ where $\omega$ is a random seed that generates a $\Phi$. It is possible to show that give $x$, with high probability, $\Phi(\omega)(x) = y$ encodes $x$ in an $\ell_2$-instance optimal fashion:
   **Equation:**

   $$\| x - x \|_{\ell_2} \leq 2\sigma_k(x)_{\ell_2}$$

   for $k \leq c_0 \frac{n}{(\log N/n)^{5/2}}$. Thus, by generating many such matrices we can recover any $x$ with high probability.

**Encoding signals**

Another practical problem is that of encoding the measurements $y$. In a real system these measurements must be quantized. This problem was addressed by Candes, Romberg, and Tao in their paper Stable Signal Recovery from Incomplete and Inaccurate Measurements. They prove that if $y$ is quantized to $y$, and if $x \in U(\ell_p)$ for $p \leq 1$, then we get optimal performance in terms the number of bits required for a given accuracy. Notice that their result applies only to the case where $p \leq 1$. One might expect that this argument could be extended to $p$ between 1 and 2, but a warning is in order at this stage:

Fix $1 < p \leq 2$. Then there exist $\Phi$ and $\Delta$ satisfying
**Equation:**

$$\| \ x - \Delta\Phi\big(x\big) \ \|_{\ell_p} \leq C_0 \sigma_k\big(x\big)_{\ell_p}$$

if
**Equation:**

$$k \leq c_0 N^{\frac{2-2/p}{1-2/p}} \left( \frac{n}{\log \ N/n} \right)^{\frac{p}{2-p}}.$$

Furthermore, this range of k is the best possible (save for the log term).

Examples:

- $p = 1$, we get our original results
- $p = 2$, we do not get instance optimal for $k = 1$ unless $n \approx N$
- $p = \frac{3}{2}$, we only get instance optimal if $k \leq c_0 N^{-2} \left( \frac{n}{\log N/n} \right)^3$